**ASAP data write-up**

The data consists of eight essay sets with between 700-2000 essays in each set. The data for each essay set consists of the essays themselves, at least two raters’ scores for each essay (the score ranges are different across the essay sets) and a resolved score. Essay set 2 is rated on two domains (the first relating to the accomplishment of the task and the second being ‘Language Conventions’), and essay sets 7 and 8 have score breakdowns by trait (content, organization, voice etc.).

For the purposes of this analysis, only one rater’s score (rater1) in one domain will be taken to be the target score for each essay. Some overall statistics for the entire data are given below. Token count metrics are obtained after using NLTK’s tokenizer.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Prompt** | **Average token count** | **Minimum token count** | **Maximum token count** | **Standard deviation of token count** | **Number of samples** | **Number of unique tokens** |
| **Prompt 1** | 417.551318 | 10 | 956 | 140.373165 | 1783 | 17840 |
| **Prompt 2** | 427.138333 | 34 | 1173 | 176.742821 | 1800 | 16632 |
| **Prompt 3** | 123.488992 | 10 | 424 | 61.810026 | 1726 | 7545 |
| **Prompt 4** | 105.172784 | 2 | 413 | 59.198453 | 1771 | 5794 |
| **Prompt 5** | 140.155679 | 5 | 486 | 67.684453 | 1805 | 5851 |
| **Prompt 6** | 171.580000 | 4 | 509 | 63.737023 | 1800 | 6401 |
| **Prompt 7** | 199.732907 | 6 | 776 | 106.138027 | 1565 | 11724 |
| **Prompt 8** | 699.735664 | 36 | 1204 | 241.433299 | 715 | 13571 |

The data is split into train, test and dev sets (60-20-20) for every prompt. The distribution of samples by score point in each of the splits is made to correspond to the overall distribution of samples by score point. Score points with less than four samples are removed from the data to allow for the distributions to match across splits. The distribution of samples by score point, by train-test-dev split, for each prompt is given in the following tables. The “disagreement rate” is the percentage of scores that differ between rater1 and rater2.

**Prompt 1:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Score point 1.0** | **Score point 2.0** | **Score point 3.0** | **Score point 4.0** | **Score point 5.0** | **Score point 6.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 6 | 17 | 118 | 552 | 304 | 72 | 1069 | 34.518241 |
| **Dev** | 2 | 5 | 39 | 185 | 102 | 24 | 357 | 34.453782 |
| **Test** | 2 | 6 | 39 | 185 | 101 | 24 | 357 | 35.294118 |

**Prompt 2:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Score point 1.0** | **Score point 2.0** | **Score point 3.0** | **Score point 4.0** | **Score point 5.0** | **Score point 6.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 14 | 92 | 458 | 467 | 45 | 4 | 1080 | 23.333333 |
| **Dev** | 5 | 31 | 152 | 155 | 15 | 2 | 360 | 17.777778 |
| **Test** | 5 | 30 | 153 | 156 | 15 | 1 | 360 | 20.555556 |

**Prompt 3:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Score point 0.0** | **Score point 1.0** | **Score point 2.0** | **Score point 3.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 25 | 408 | 413 | 189 | 1035 | 24.830918 |
| **Dev** | 8 | 136 | 138 | 63 | 345 | 25.217391 |
| **Test** | 8 | 136 | 139 | 63 | 346 | 26.011561 |

**Prompt 4:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Score point 0.0** | **Score point 1.0** | **Score point 2.0** | **Score point 3.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 188 | 454 | 313 | 107 | 1062 | 22.881356 |
| **Dev** | 63 | 151 | 104 | 36 | 354 | 22.033898 |
| **Test** | 63 | 151 | 105 | 36 | 355 | 23.380282 |

**Prompt 5:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Score point 0.0** | **Score point 1.0** | **Score point 2.0** | **Score point 3.0** | **Score point 4.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 31 | 226 | 417 | 291 | 118 | 1083 | 42.012927 |
| **Dev** | 10 | 75 | 139 | 98 | 39 | 361 | 42.659280 |
| **Test** | 10 | 76 | 139 | 97 | 39 | 361 | 41.551247 |

**Prompt 6:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Score point 0.0** | **Score point 1.0** | **Score point 2.0** | **Score point 3.0** | **Score point 4.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 33 | 127 | 278 | 484 | 158 | 1080 | 38.703704 |
| **Dev** | 11 | 43 | 92 | 162 | 52 | 360 | 39.444444 |
| **Test** | 11 | 43 | 92 | 161 | 53 | 360 | 33.055556 |

**Prompt 7:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **SP 2.0** | **SP3.0** | **SP 4.0** | **SP 5.0** | **SP 6.0** | **SP 7.0** | **SP 8.0** | **SP 9.0** | **SP 10.0** | **SP 11.0** | **SP 12.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 4 | 25 | 57 | 73 | 83 | 76 | 252 | 124 | 83 | 58 | 104 | 939 | 70.607029 |
| **Dev** | 1 | 9 | 19 | 24 | 28 | 26 | 84 | 41 | 27 | 19 | 35 | 313 | 71.246006 |
| **Test** | 1 | 9 | 19 | 24 | 28 | 25 | 84 | 41 | 28 | 19 | 35 | 313 | 70.926518 |

Score points removed: 0.0, 1.0

**Prompt 8:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **SP 10.0** | **SP 11.0** | **SP 12.0** | **SP 13.0** | **SP 14.0** | **SP 15.0** | **SP 16.0** | **SP 17.0** | **SP 18.0** | **SP 19.0** | **SP 20.0** | **SP 21.0** | **SP 22.0** | **SP 23.0** | **SP 24.0** | **SP 25.0** | **Total** | **Disagreement rate (%)** |
| **Train** | 6 | 4 | 5 | 8 | 5 | 68 | 28 | 44 | 43 | 18 | 147 | 14 | 8 | 11 | 4 | 16 | 429 | 71.561772 |
| **Dev** | 2 | 1 | 1 | 2 | 2 | 23 | 9 | 15 | 15 | 6 | 49 | 4 | 3 | 4 | 1 | 6 | 143 | 72.727273 |
| **Test** | 2 | 1 | 2 | 3 | 2 | 22 | 9 | 15 | 15 | 6 | 49 | 4 | 3 | 4 | 1 | 5 | 143 | 73.426573 |

Score points removed: 5.0, 26.0, 27.0, 28.0, 30.0